

van leraren nogal wisselend te zijn. Bestaande opvattingen van leraren over onderwijzen blijken vaak zeer hardnekkig te zijn. Deze opvattingen passen vaak meer bij een transmissie-model van het onderwijsleerproces dan bij een constructivistische benadering.

In de papers lag overwegend de nadruk op de kennis en opvattingen van docenten als aangrijpingspunt voor veranderingen. Het in de praktijk brengen van nieuwe modellen van onderwijzen, vraagt echter ook om de verwerving van nieuwe vaardigheden. Op de conferentie was nauwelijks aandacht voor gerichte vaardigheidstraining als component van een programma voor professionele ondersteuning. Evenmin werd aandacht geschonken aan de rol die curriculummaterialen kunnen spelen in een dergelijk traject. Wellicht dat een meer integrale benadering de effectiviteit en efficiency van dergelijke ondersteuningsprogramma's kan vergroten.

Schooleffectiviteitsonderzoek (H. de Vos, Universiteit Twente)

Alhoewel er op de AERA in 1995 zo'n 1400 papersessies waren, werden er maar drie gekenmerkt als schooleffectiviteit-sessies. Eén ging over de relatie van schooleffectiviteitsonderzoek en schoolverbetering, één over onderwijs indicatoren en de laatste ging over voortgang van schooleffectiviteitsonderzoek. De sessies kenmerkten zich door hun internationale karakter: de presentaties werden niet alle door Amerikanen verzorgd.

Aangezien mijn interesse voornamelijk methodologisch was getint, zal ik met name de presentaties eruit lichten, die aan mijn interesse tegemoet komen.

Een geheel Nederlandse sessie was verroosterd op donderdagmorgen: The evaluation of the Dutch Educational Priority Program. Het publiek bestond grotendeels uit Nederlanders. Het onderwerp was, voor zover dat nog niet duidelijk was uit de titel, de evaluatie van zes jaren OVB. De participanten aan deze sessie waren Van Tilborg en Walraven van SVO, Mulder van het ITS en Van der Werf van het GION. Alhoewel blijkt dat de prestaties op de OVB-scholen verbeterd zijn, is het moeilijk te zeggen of dit ook door het OVB is bewerk-

stelligd. De oorzaak zou ook kunnen liggen in het feit dat de leerlingen zes jaar langer in Nederland wonen. Welk deel van de verbetering wordt verklaard door factoren binnen de school, factoren buiten de school en OVB is nog onduidelijk. De discussiant Slavin, van de John Hopkins University, was erg enthousiast over de OVB, mede omdat het Nederlandse onderwijsbeleid ten aanzien van specifieke doelgroepen soms lijnrecht tegenover de uitkomsten van het gedecentraliseerde onderwijs-financieringsstelsel van de Amerikaanse staten staat: (relatief) arme staten besteden minder geld aan de compensatie van achterstanden van gedepriveerde groepen leerlingen.

Ook donderdagochtend (om 8.15 uur!) was er een sessie over internationaal gebruik van onderwijs-indicatoren: International uses of quality indicators. De toevoeging "internationaal" was zeker op zijn plaats; bijdragen aan deze sessie kwamen uit Canada, Engeland, Frankrijk, Nederland en de Verenigde Staten. Onderwijsindicatoren worden gebruikt om de kwaliteit van onderwijs en scholen te beschrijven. De papers belichtten zienswijzen omtrent ontwikkeling en optimaal gebruik van deze indicatoren. In Canada, aldus McEwen, bepalen de provincies het onderwijs en dus ook het gebruik van indicatoren. Zij stelde een aantal verbeteringen voor op provinciaal en nationaal niveau, zoals het gebruik van indicatoren voor vergelijkingen tussen provincies (een soort benchmarking?) en voor het optimaliseren van schoolbegeleiding. Uit Engeland kwam een bijdrage van Fitz-Gibbon. Zij vroeg zich af of het systeem van indicatoren dat wordt gebruikt in de UK, wel een beeld geeft van de kwaliteit van onderwijs. Het indicatorensysteem wordt met name bepaald door examenresultaten en resultaten van onderwijsinspectie. De kwaliteit van het indicatorensysteem wordt momenteel verhoogd door de toevoeging van 'value-added' maten. Creemers (Nederland) ging in op de bijdrage van schooleffectiviteitsonderzoek aan de ontwikkeling en het gebruik van onderwijsindicatoren. Hij stelde dat onderzoek, zoals IEA (International Association for the Evaluation of Educational Achievement) en ISERP (The International School Effectiveness Research Program), de ontwikkeling ondersteunen. Het is echter wel zaak, aldus

Creemers, dat indicatoren internationaal vergelijkbaar zijn, dus onafhankelijk van het onderwijsstelsel. Bonnet uit Frankrijk hield een pleidooi voor het monitoren van het onderwijs met behulp van indicatoren. Hij waarschuwde dat moet worden opgepast dat indicatoren geen politieke instrumenten worden.

Op de vrijdagmiddag was er een sessie over internationale voortgang in schooleffectiviteitsonderzoek georganiseerd door ISERP, onder voorzitterschap van Reynolds. Tymms (UK) ging in zijn presentatie 'Monitoring effectiveness: Examples of the evaluation of complex systems' ver door te beweren dat schooleffectiviteitsonderzoek er nooit in zal slagen zodanige kennis te vergaren dat scholen in het algemeen kunnen worden verbeterd, omdat wat er in scholen en klassen gebeurt onvoorspelbaar, complex en chaotisch is. Hij hield een pleidooi voor gebruik van onderwijsindicatoren, die meer inzicht kunnen geven in wat er nu eigenlijk gebeurt. In de paperpresentatie over 'Two year quantitative findings from ISERP' (o.a. Creemers) werd ingegaan op het gegeven dat de leerlingvariabelen (SES, IQ, ethniciteit) het grootste deel van de wiskundeprestaties van leerlingen verklaren in internationaal vergelijkend onderzoek (net als in het meeste onderzoek). Alhoewel in de meeste landen ook scholenfactoren van invloed zijn op de leerlingprestaties is dat in Nederland (en ook in Taiwan) niet het geval.

Measurement and Research Methodology (H. J. Vos, Universiteit Twente)

Het aandachtsgebied 'Measurement and Research Methodology' was ruim vertegenwoordigd in de vorm van thematische papersessies, (invited) symposia waarvan 36 in samenwerking met de National Council on Measurement in Education (NCME), posterpresentaties en pre-conference workshops.

Een van de belangrijkste trends betrof het adaptief toetsen van vaardigheden. Computerized Adaptive Testing (CAT) wordt steeds meer gebruikt in de praktijk en lijkt de toetsvorm van de toekomst te worden. Het belangrijkste voordeel van deze toetsvorm in vergelijking met conventionele potlood-en-papier (P&P) toetsen is dat volstaan kan worden met kortere toet-

sen bij gelijkblijvende of zelfs hogere nauwkeurigheid van de geschatte vaardigheid. Er is echter nog veel psychometrisch onderzoek te doen op dit gebied, zowel voor 'fine-tuning' van de theorie als voor het oplossen van praktische problemen bij de implementatie. Aan beide aspecten van CAT werd in verschillende papersessies en symposia ruimschoots aandacht besteed.

Veerkamp en Chang stelden beiden nieuwe criteria voor om items te selecteren in een adaptieve toetsprocedure gebruikmakend van item response theorie (IRT) modellen. Het belangrijkste bezwaar bij de meeste huidige criteria is dat ze geen rekening houden met de onzekerheid van de geschatte vaardigheid. Terwijl Veerkamp een gewogen gemiddelde van de gebruikelijke informatiefunctie voorstelde, stelde Chang daarentegen een 'global information approach' voor om de gesignaleerde bezwaren bij de huidige item selectiecriteria te omzeilen.

In de presentatie van Rocklin, Vispoel en Wang over de mogelijkheid tot manipulatie van het CAT-algoritme als item review mogelijk is door de 'test-taker' (Self-Adapted Testing) kwam naar voren dat het erg riskant is voor de test-taker om op deze manier te trachten onterecht een hogere score te behalen. In theorie kan een test-taker door er voor te zorgen alleen erg gemakkelijke items voorgelegd te krijgen en die vervolgens allemaal goed te beantwoorden kunstmatig een erg hoge score verkrijgen. Als echter enkele vragen fout beantwoord worden zal zijn of haar score te laag uitvallen.

Lunz en DeVille deden verslag van een studie waarin toetsen geconstrueerd via de computer met een adaptieve selectieprocedure werden vergeleken met toetsen die handmatig werden geconstrueerd en op papier werden afgenomen. Hierbij werd met name gekeken naar de validiteit van de geselecteerde items, de volgorde van de items, de statistische eigenschappen van de toetsen en in hoeverre de geselecteerde items voldeden aan de toetsspecificaties. Uit deze studie bleek dat beide methoden voor toetsconstructie vergelijkbaar zijn voor wat betreft validiteit en statistische eigenschappen en dat beide methodes voldeden aan de toetsspecificaties.

Davey en Parshall keken naar exposure